Great Learning – Artificial Intelligence and Machine Learning

Credit Risk Modelling

Project report by – Arunangshu Podder, Garima Bajpai, Madhvi Gaur, Nikhil Gupta

---

Problem Statement

The main objective of this project is to:

1) identify the factors that influence a loan getting repaid successfully or closed on account of default,
2) based on those factors, build a model that can predict whether a borrower will eventually default on his loan or not, and
3) build separate model that will, in the event of a default, predict the net impact of the default on the lender.

Through this project, we have aimed to help individual lenders or lending institutions minimize the risk of losing money due to bad loans.
Such loan default prediction and loss calculation scenarios are part of *Credit Risk Modelling* and form a classic data science problem. This also provided an opportunity to apply various supervised and unsupervised methodologies of data science to solve the business problem.
As a part of our prediction problem, we have built a model which on the basis of the current attributes of the dataset, will predict whether an ongoing loan will eventually default or not. If a loan defaults, we will use the second model, built as part of our prediction, which will predict how much the bank is expected to lose. We have expressed the same as Loss Given Default or LGD which is defined as the proportion of the total exposure when the borrower defaults.
Through this problem we have also analysed the different attributes present in a loan application that contribute to a default loan.
We have built our business models using a conservative approach using rigorous evaluations since rejecting too many prospective borrowers can hamper the business.

Data

For this proposed project, we have looked into data published by various Peer to Peer (P2P) lending platforms since it is very difficult to obtain real bank data. P2P platforms, such as Prosper, Lending Club, Kiva, Fynanz provide various online services, thus enabling individuals and small businesses to get hassle-free loans from interested lenders. Amongst them, Lending Club is the largest P2P lending platform in the world.
The dataset released by Lending Club is currently available in Kaggle (https://www.kaggle.com/wordsforthewise/lending-club), containing 2 million+ loan records issued between the years 2007 and 2018.
The dataset has 151 variables which include information such as borrower's credit history, personal information (e.g. annual income, years of employment, zip code), loan information (e.g. description, type interest rates, grade), current loan status, etc. It also contains some variables which define how the settlement will be done for defaulted loans. Such information is a knowledge of the future and need to be handled accordingly.

Process overview

- Salient Features of the data:

As part of the business solving process, we did a thorough analysis of the dataset. [5] tells us about the lending process of Lending Club. A prospective borrower applies, reporting information about himself, his finances, and his need for a loan. Lending Club checks the borrower's score and report, then assigns him a risk subgrade and a corresponding interest rate. If the borrower accepts the interest rate, the loan is listed on the Lending Club website. Prospective lenders can browse the loans listed on the website and agree to fund a portion of the loan. The loan is issued through a bank affiliated with Lending Club, so these loans are legally the same as traditional unsecured loans. However, Lending Club does not make loans until borrowers have agreed to fund the entirety of the loan. All issued loans are thus immediately securitized. Lending Club may choose to verify the borrower's income while the loan is in funding but does not delay the listing of the loan to verify income. If the borrower's income cannot be verified, the loan is removed from the site. If the loan becomes fully funded before the borrower's income can be verified, the loan is issued without verification.

Some of the important variables in the dataset are listed in the below table.

Variable Name	Data Type	Description
addr_state	Categorical	The state provided by the borrower in the loan application.
application_type	Categorical	Indicates whether the loan is an individual application or a joint application with two co-borrowers.
loan_amnt	Numeric	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
funded_amnt	Numeric	The total amount committed to that loan at that point in time.
installment	Numeric	The monthly payment owed by the borrower if the loan originates.
int_rate	Numeric	Interest Rate on the loan.
issue_d	Date	The month and year when the loan was funded.
grade	Categorical	LC assigned loan grade.
sub_grade	Categorical	LC assigned loan subgrade.
verification_status	Categorical	Indicates if income was verified by LC, not verified, or if the income source was verified.
term	Categorical	The number of payments on the loan. Values are in months and can be either 36 or 60.
purpose	Categorical	A category provided by the borrower for the loan request.
annual_inc	Numeric	The self-reported annual income provided by the borrower during registration.
emp_length	Categorical	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
home_ownership	Categorical	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
acc_now_delinq	Numeric	The number of accounts on which the borrower is now delinquent.
dti	Numeric	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
fico_range_high	Numeric	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	Numeric	The lower boundary range the borrower's FICO at loan origination belongs to.
pub_rec_bankruptcies	Numeric	Number of public record bankruptcies.
open_acc	Numeric	The number of open credit lines in the borrower's credit file.
total_pymnt	Numeric	The total number of credit lines currently in the borrower's credit file.

Table 1

- loan_status (current status of the loan)

The feature **"**loan_status" has the following values as:

1. Current: Active Loan.
2. Fully Paid: The full principal with interest rates is paid back.
3. Charged Off: The borrower defaulted on the loan and the loan will never be paid back in full amount.
4. In Grace Period: Payment of installment is delayed by 1 to 15 days.
5. Late (16–30 days): Payment of installment is delayed by 16 to 30 days.
6. Late (31–120 days): Payment of installment is delayed by 31 to 120 days.
7. Default: Payment of installment is delayed by more than 120 days.

We find in figure 1, that our data comprises of approx. 19.6% of Charged Off loan samples which we can use for our model training purpose, and then use the same model on the currently active loan samples that have been marked as Bad Loans by our default classification model.

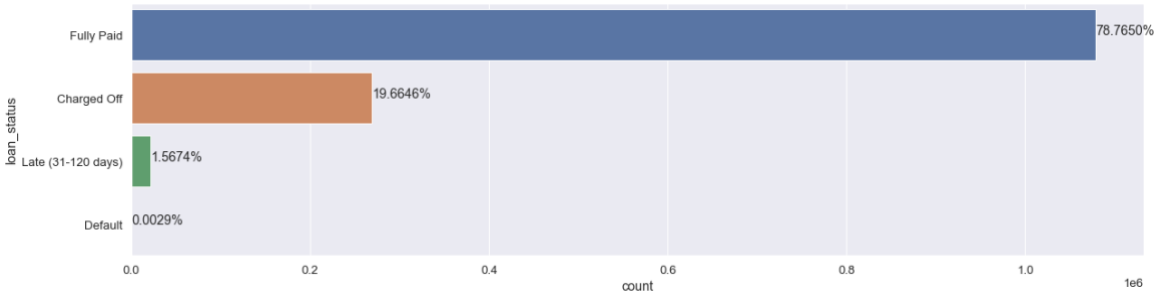
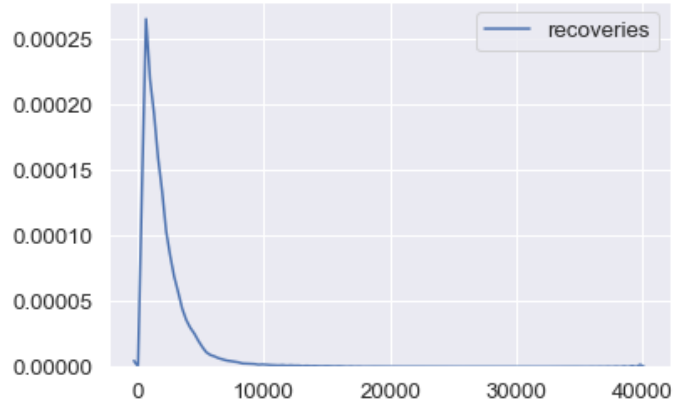


Fig 1



Also, the KDE plot in figure 2 shows the distribution of the variable "recoveries". Here, we find the variable to be highly skewed and a large chunk of the data to be around 0.

Fig 2

- **issue_d** (Interest Rate on the loan)

The loans present in the dataset are issued between 2007 and 2018. The plot in Figure 3 shows that there has been a steady rise in the number of loans issued by Lending Club. Studies show that there has been a massive rise in peer-to-peer lending platforms for micro financing over the recent years and the graph in figure 3 proves the same.

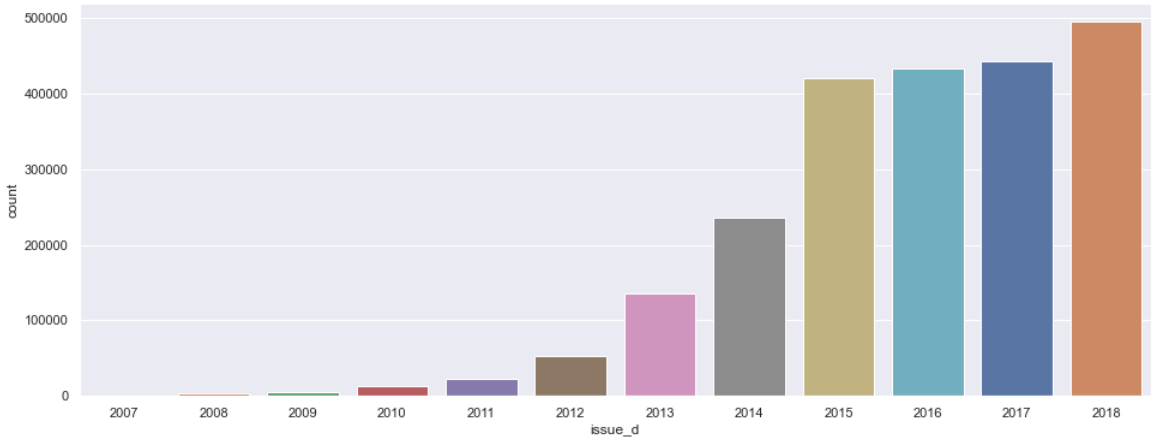


Fig 3

- **term** (the tenure on the loan)

When a borrower applies for a loan, Lending Club gives 2 options for loan tenure, viz, 36-months and 60 months. From Figure 4, we find that *when Lending Club started in 2007, it started with only 36-month loan terms. From 2010 onwards, it introduced the 60-month loan tenure option* (as shown in Figure 4).

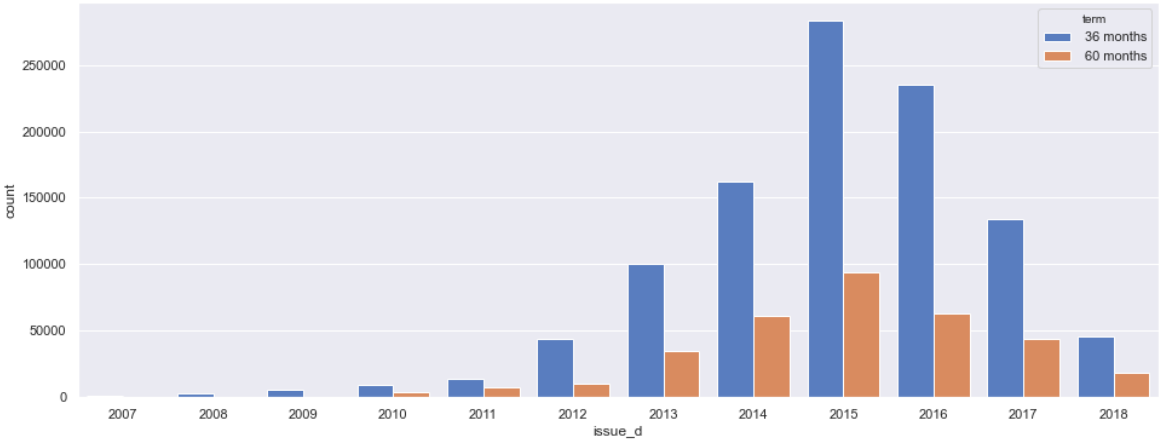


Fig 4

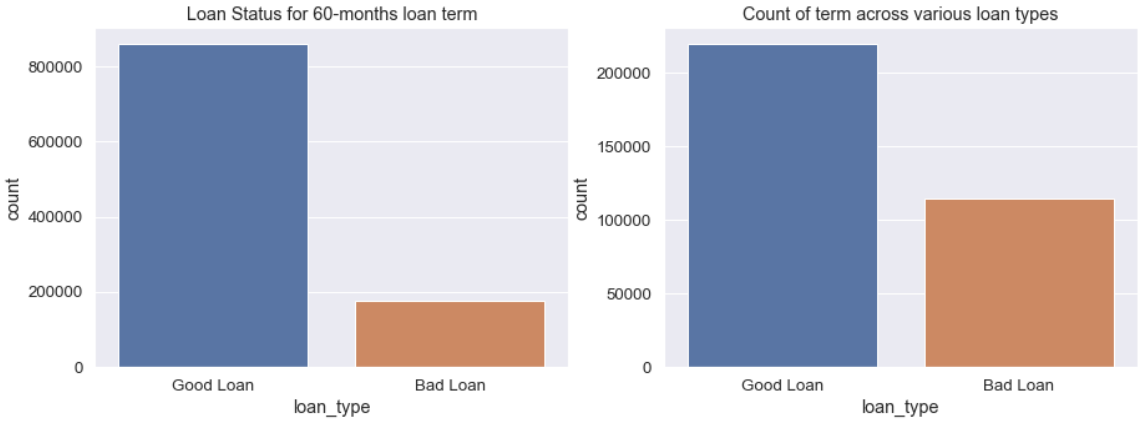


Fig 5

The plots in Figure 5 show that there is a big difference between the count of good and bad loans for 36-months loan tenures compared to that of 60-months. This shows that *a loan has more probability to default if it has a longer tenure.* However, this needs further investigation to check why loans with longer tenure are more prone to default.

- **loan_amnt** (the listed amount of the loan applied for by the borrower)
- **funded_amnt** (the total amount committed to that loan at that point in time)
- **funded_amnt_inv** (the total amount committed by investors at that point in time.)

Next, we look at loan amount. There are 3 variables in the dataset **"loan_amnt"** (the listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value), **"funded_amnt"** (the total amount committed to that loan at that point in time) and **"funded_amnt_inv"** (the total amount committed by investors for that loan at that point in time).

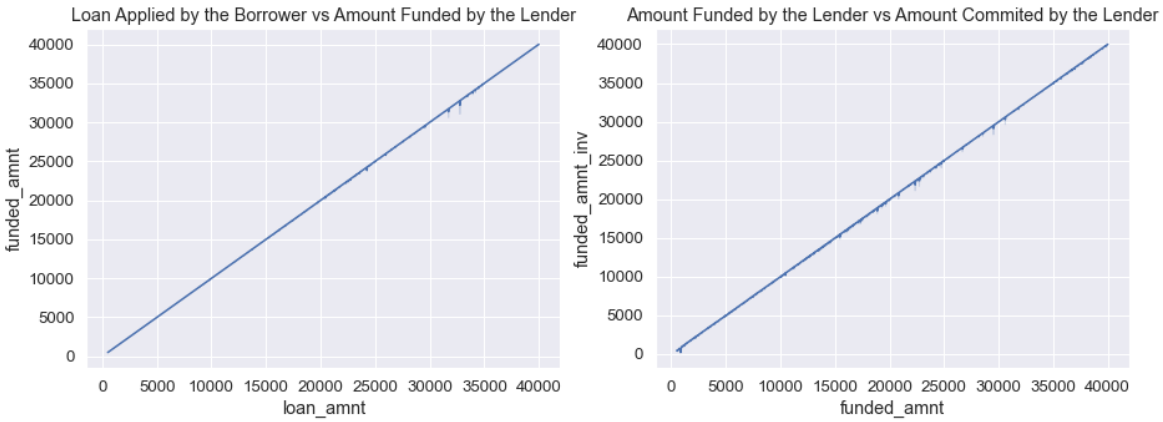


Fig 6

The line plots in figure 6 show a linear relation between the 3 variables which means that *the loan amount the borrower is applying for, is the same amount the lender is committing to invest.* Thus, we can pick any one of them and drop the other two.

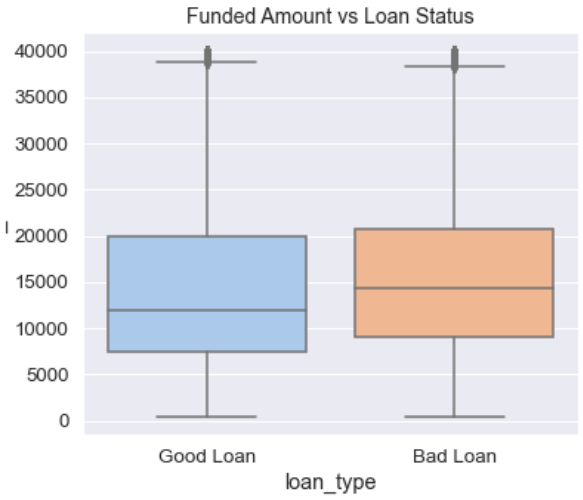
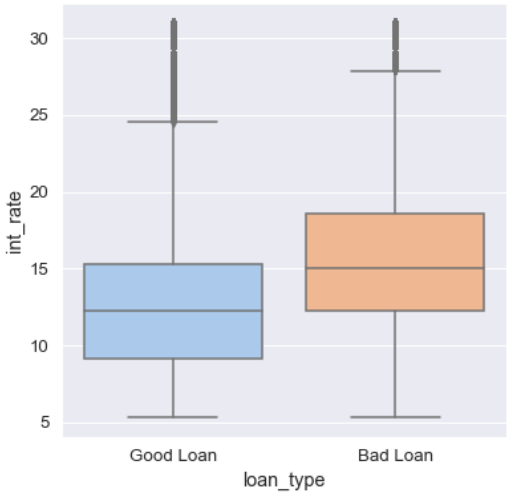


Fig 7

Here, picking **"funded_amnt"** makes more sense since it is the amount on which the borrower will be paying the installments. The box plot (figure 7) shows the mean of funded amount in bad loans is comparatively higher than good loans. *This proves that loans with higher amounts are more prone to default.*

- **int_rate** (interest rate on the loan)

Next, we focus on the variable for interest rate that the borrowers need to pay on their loans. From the box plot (figure 8), we find that the mean interest rate for defaulted loans to be much higher than for loans that ended successfully. This tells us that *loans with higher interest rate have a greater tendency to default.*



- **installments** (the monthly payment owed by the borrower if the loan originates)

The dataset provides us information about the installments that the borrower needs to pay once the loan is approved. We know that installments are dependent on factors such as loan amount/funded amount, interest rate and loan tenure. Fig 8

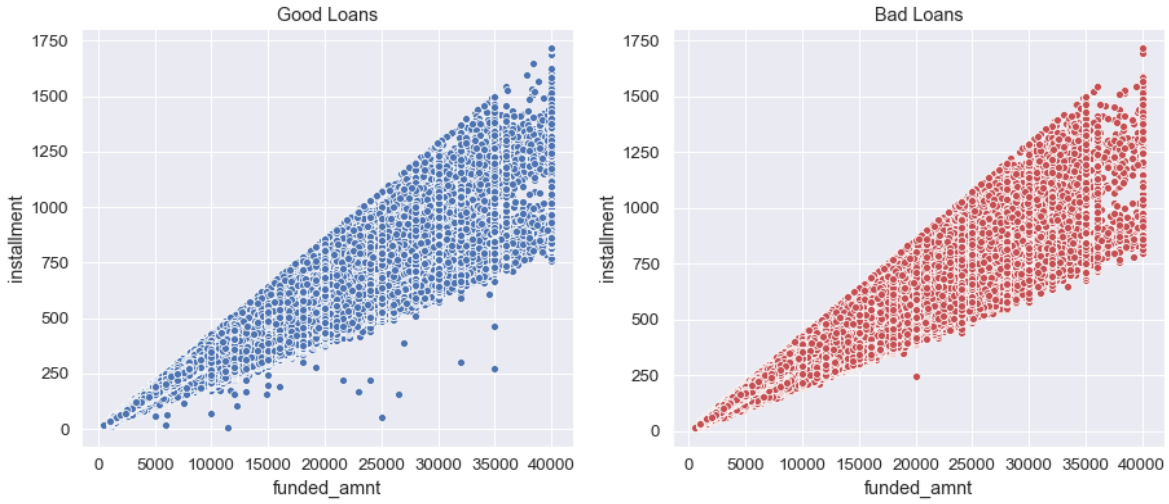


Fig 9

Figure 9 shows *the scatter plot between installment and funded amount. It clearly shows a strong correlation between the two variables.* Hence, we drop the variable for installment from our model.

- **grade** (LC assigned loan grade)
- **sub_grade** (LC assigned loan subgrade)

Now, we focus on the variables for grades. Lending Club has a system of assigning a grade and a sub-grade to every lender based on their personal details entered by them like annual income as well as on their credit report information like FICO scores.

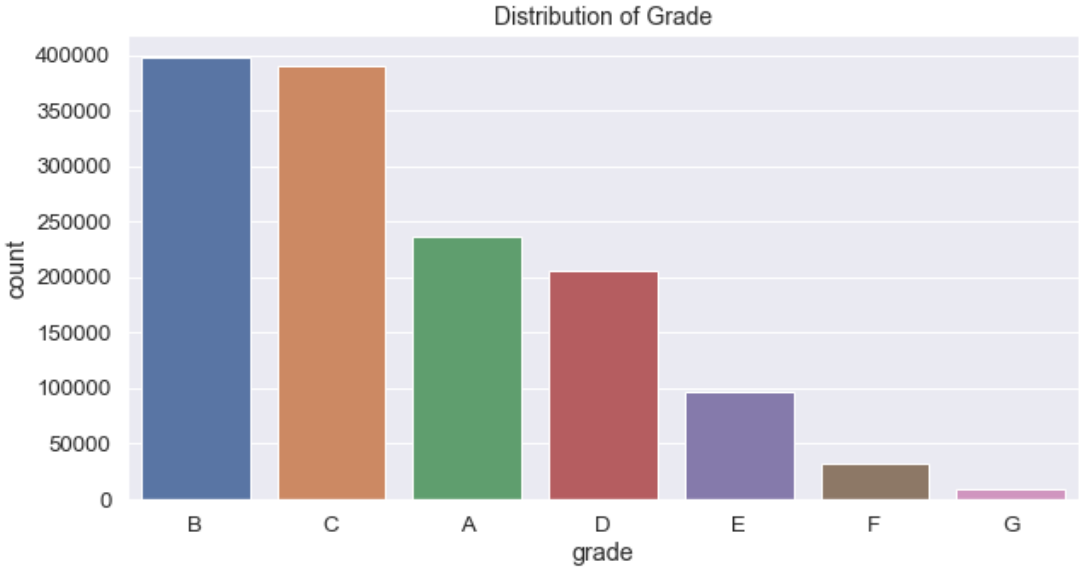


Fig 10

Grades are issued between A and G, A being the safest and G the riskiest. Figure 10 gives us a clear picture of the grades. It seems there is a big share of lenders falling in the category of B and C, which are considered to be comparatively safer (good and average credit history). A is given to the customers who have the best credit history and of course the count of such customers will be less. As we move to riskier grades, the count of accepted loans gradually decrease.
Each grade again has 5 more divisions or sub-grades. (shown in figure 11).

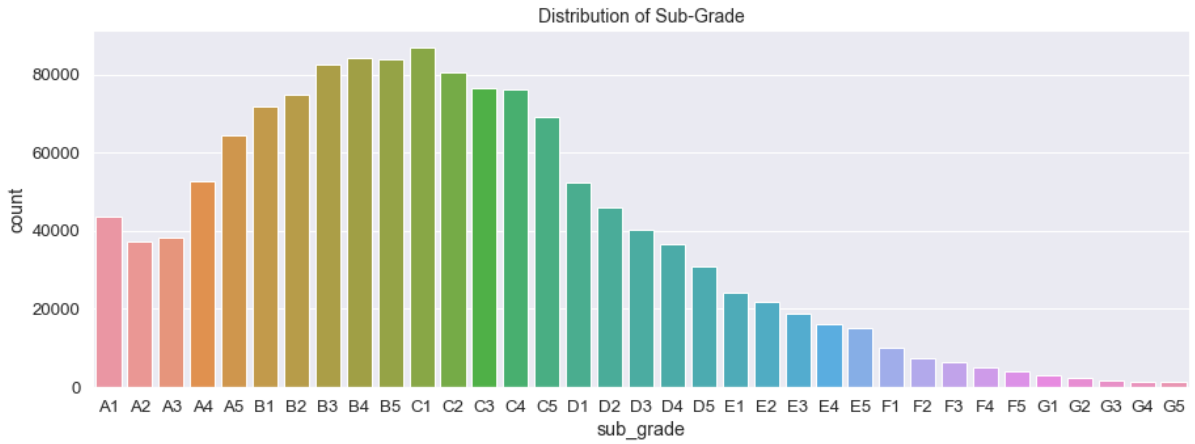


Fig 11

Since sub-grade is a more specialized division of grade, we can keep sub-grade and drop off grade.

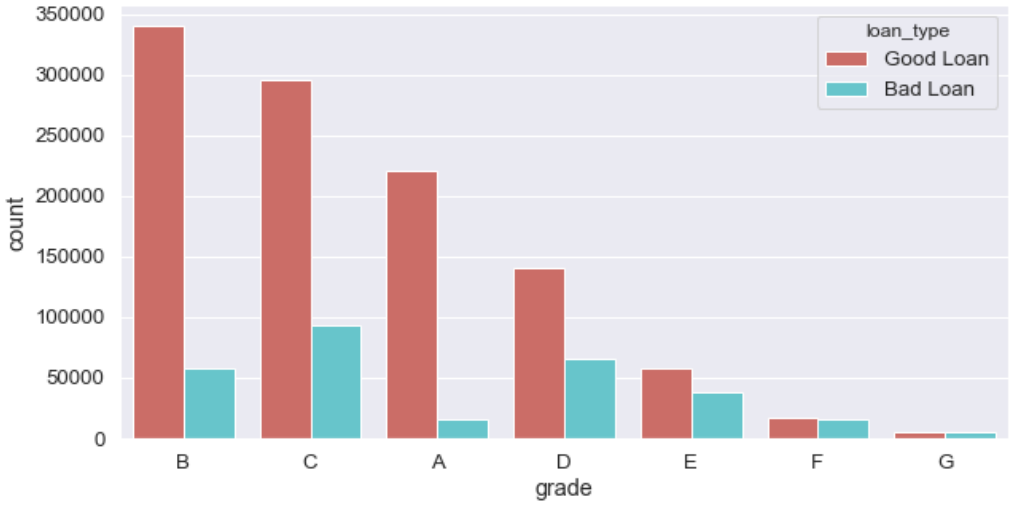


Fig 12

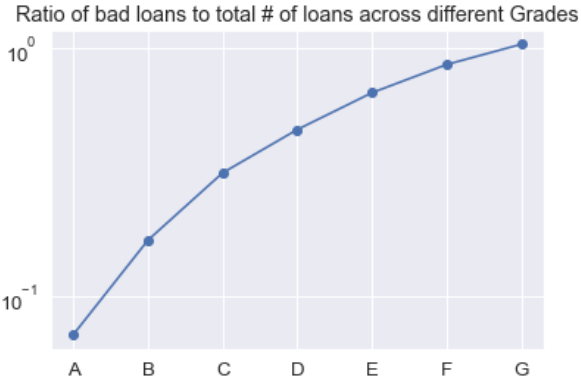


Figure 12 shows the distribution of Good and Bad loans across different grades. However, since the count of loans with grade as B and C are higher, they clearly dominate the plot. Hence, we need to take a look at the plot in figure 13 which gives us the ratio of bad loans to the total number of loans. *This clearly shows that A is the safest grade whereas G is the riskiest.*

Fig 13

In figure 14, we also get to see the mean of funded amounts / loan amount applied increases as we move towards the riskier grades. We have already seen that higher the loan amount, higher is the probability for a loan to default.

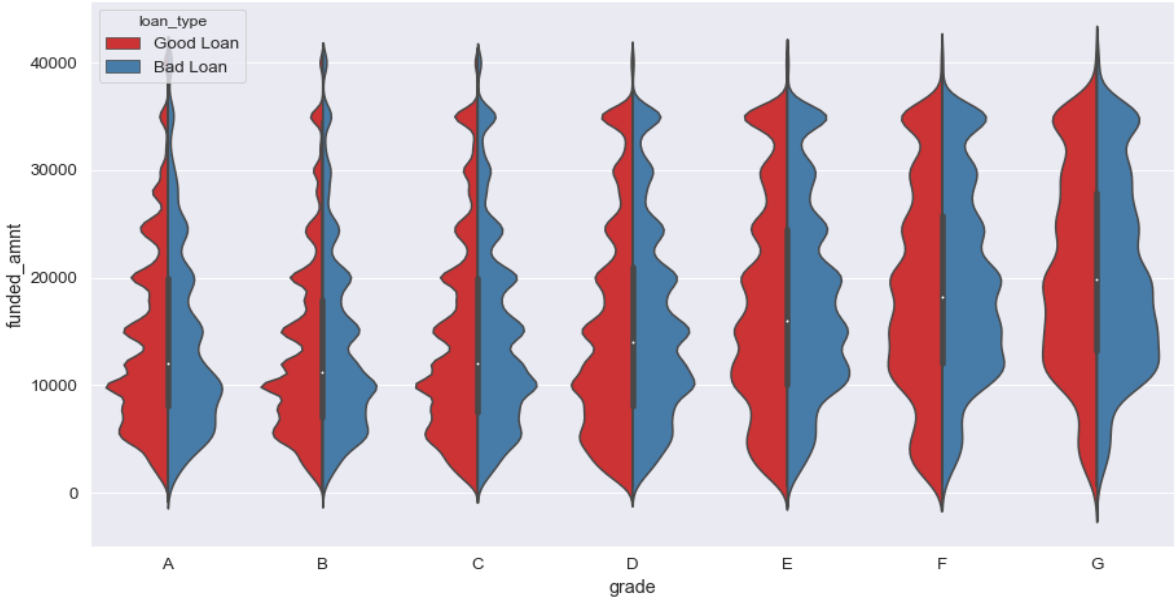


Fig 14

- **annual_inc** (the self-reported annual income provided by the borrower during registration)

Earlier, we mentioned that the grading system of LC depends on the annual income of the borrower. However, the violin plot in figure 15 does not seem to show any such relation. Although we find that borrowers having high annual income are under grade A, but for the remaining grades, we don't find any proper relation.

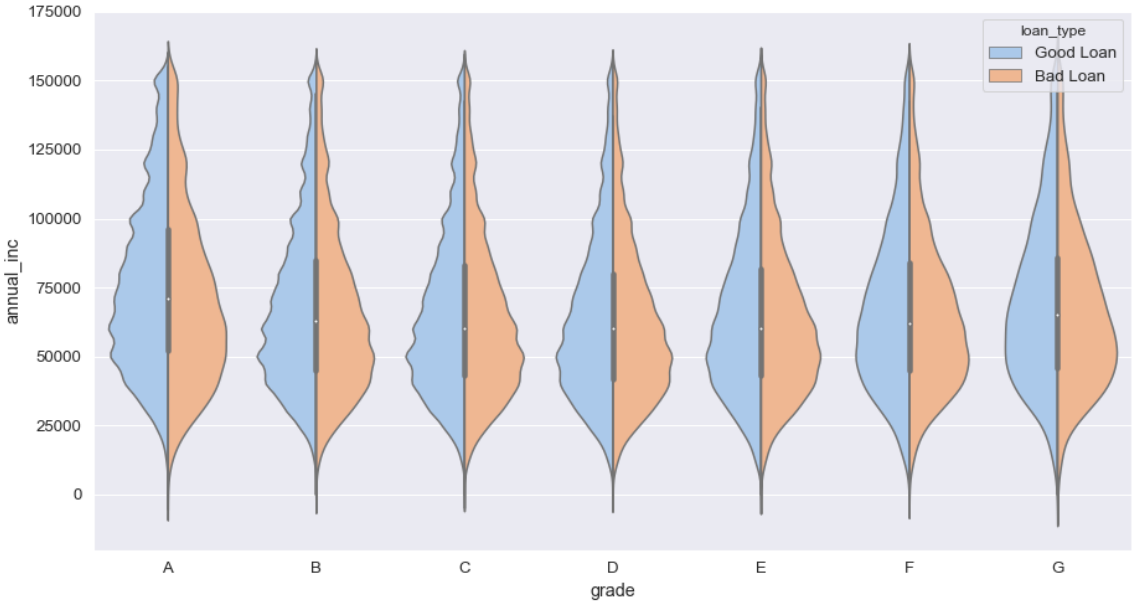


Fig 15

- **emp_length** (employment length in years.)

So, we took a look at the variable for employment length. The different values under employment length ranges from <1 year (we consider it as 0) to 10+ years. From the peaks in the graph in

figure 16, we find that people with more than 10 years of work experience tend to take more loan. Hence, the count of default is also high in that category. Apart from that, we find some peak for lenders having 1-year work experience, which indicate that there is a slight tendency for people who have just started career to default on their loans.

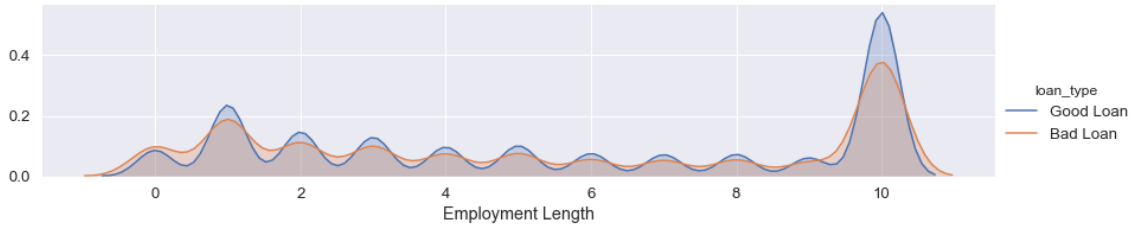


Fig 16

- **home_ownership** (the home ownership status provided by the borrower during registration or obtained from the credit report)

Another variable under personal category is for home ownership, which states whether the borrower has a owned house or not. Here, we find that those with mortgages on their homes tend to take more loans (figure 17) as well as larger amounts of loans (figure 18).

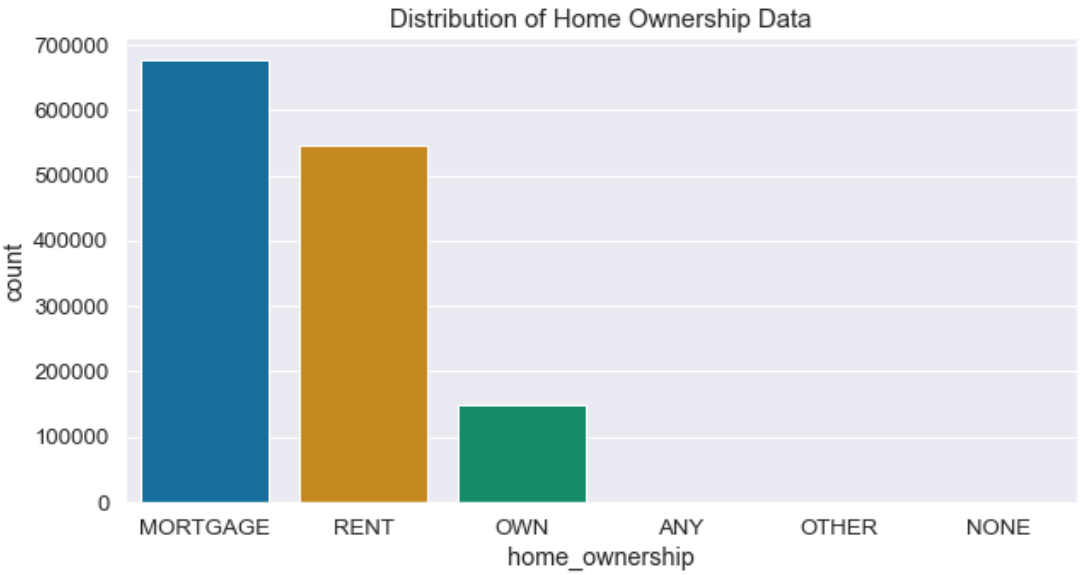


Fig 17

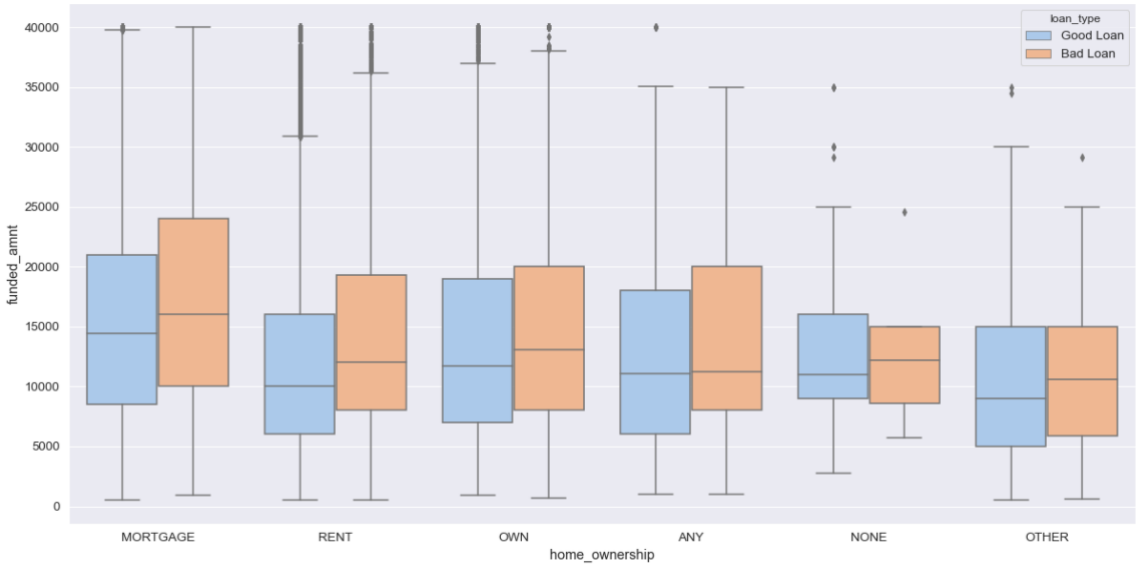


Fig18

- **addr_state** (the state provided by the borrower in the loan application)

Another variable under the personal information category in the dataset in the address variable which states the state where the user is residing. Fig 19 shows that most of the loans originate in the states of CA, TX, NY and FL whereas IA has the least number of borrowers. Because of the sheer number of loans originating from CA, the number of bad loans count is also high in CA. Thus, we can say that a new borrower coming from CA state has some probability to default.

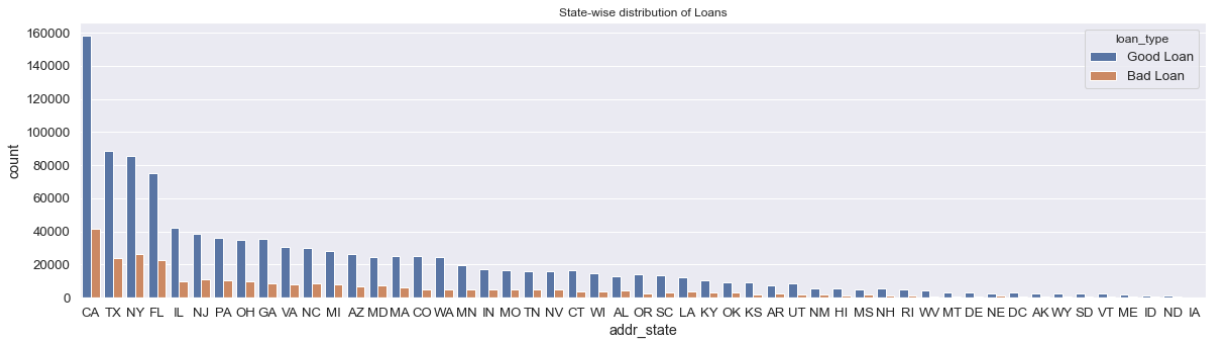


Fig 19

Since, we have the variable annual income with us, we get an opportunity to look at the average annual income of the states given in the dataset. This might help us to understand why people from states CA, TX, NY and FL are taking so much loans, whereas states like ID, ND and IA are taking so less loans.

Figure 20 shows that states like DC, CT, MD, etc have the highest average loans. However, the count of loans from these states are quite less. So, we cannot say that there is a relation between annual income and state.

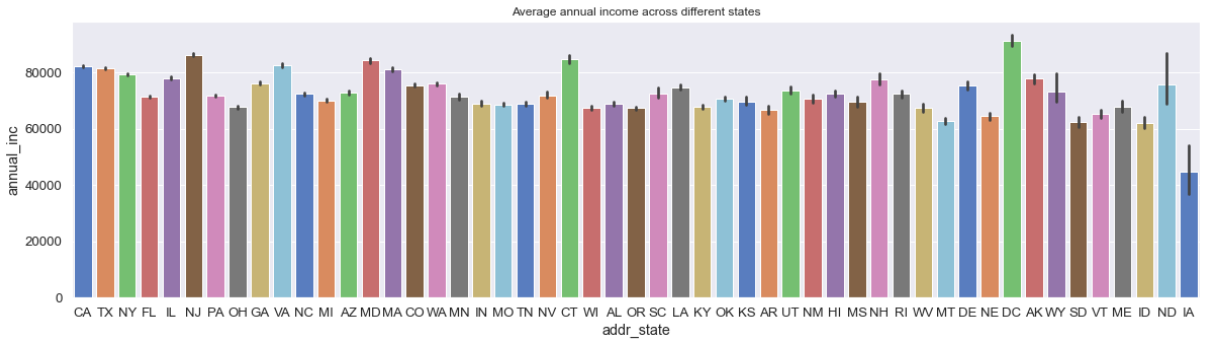


Fig 20

- **purpose** (a category provided by the borrower for the loan request)

In order to apply for a loan, lending club asks to state the purpose of the loan. So, we checked the purposes that produce the highest number of bad loans. From figure 21, we find that *most of the bad loans occur when taken for funding small businesses.*

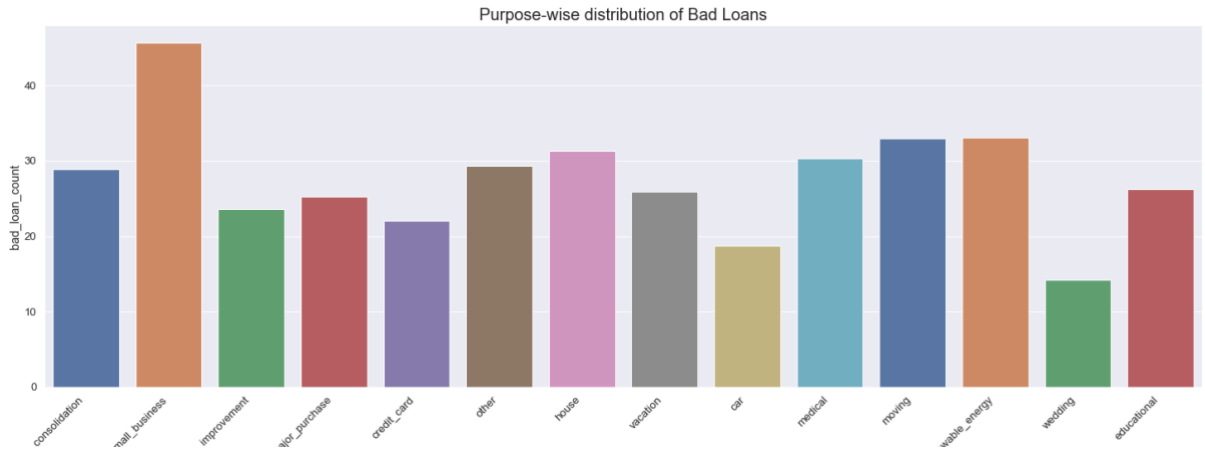
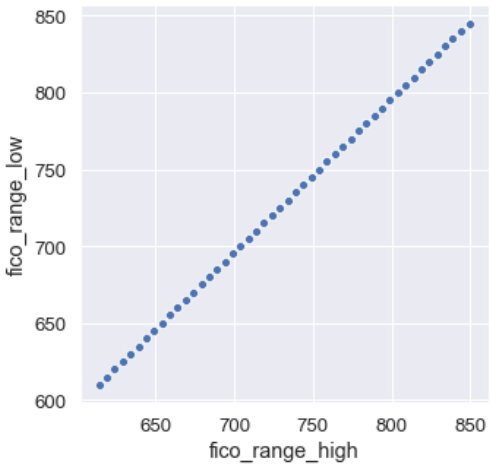


Fig 21

- **fico_range_high** (the upper boundary range of the borrower's FICO at loan origination)
- **fico_range_low** (the lower boundary range of the borrower's FICO at loan origination)

One important variable in the credit history of a borrower is FICO score. As per LC, based on this data, grades and sub-grades are assigned to the borrower's loan application. In the dataset, we find 2 variables, one which states the lower range of the borrower's FICO score and the other one that state the higher range of the score. To understand how the range differs, we plot the values on a graph (on the right). We find that the values follow a complete straight line, meaning that the difference between the 2 variables is always the same. Hence, we can drop any one of them.



Next, we check how FICO scores are affecting the grading system of LC. From the violin plot in figure 23, we find that the mean FICO score of grade A borrowers is much higher and the mean values decrease as we move towards riskier grades. This clearly proves that LC depends heavily on the FICO scores of the borrowers to assign grades.

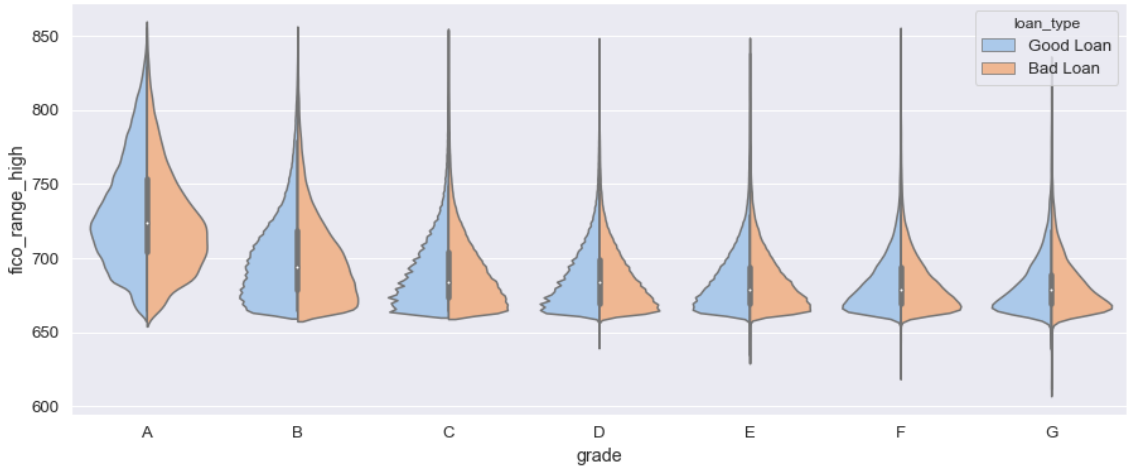


Fig 23

We also get to see how FICO score itself depends on the performance of the borrower as per his credit report. Some of the variables are listed below.

- # of charge offs within the last 12 months
- # of delinquent accounts over the last 24 months
- # of accounts currently past the 30 days due mark

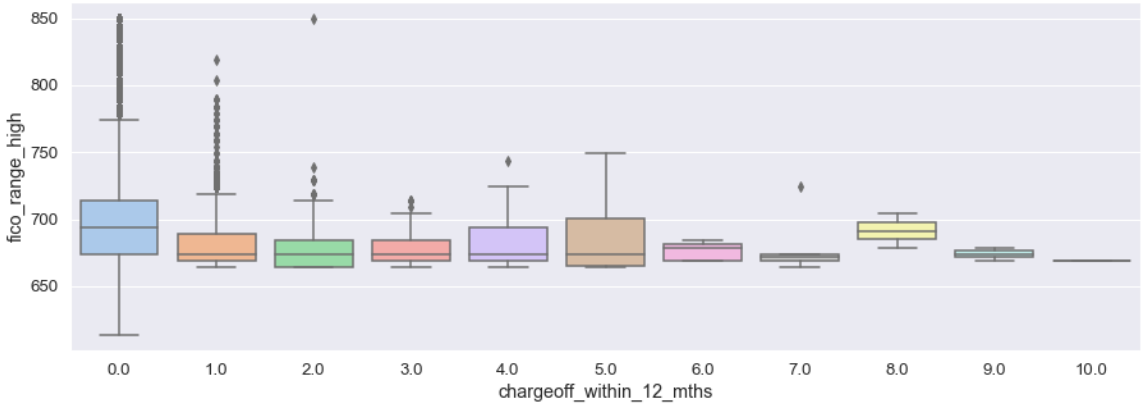


Fig 24

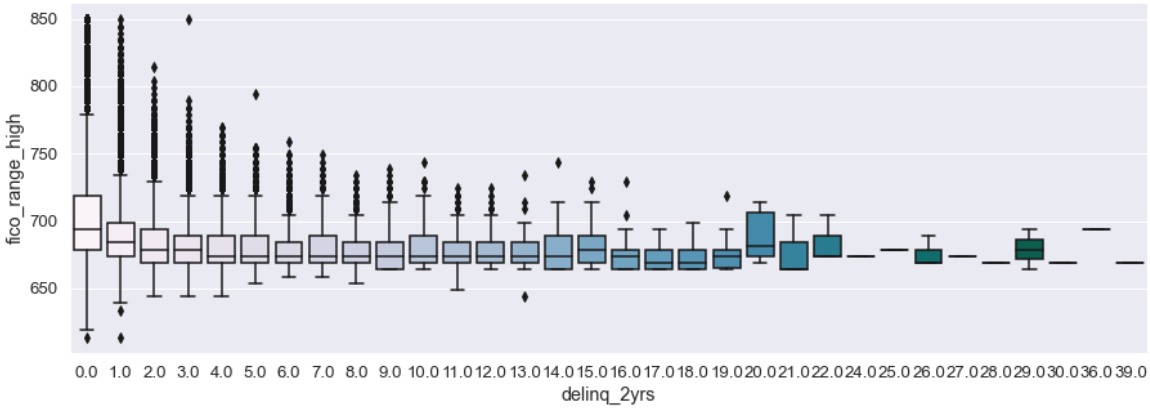


Fig 25

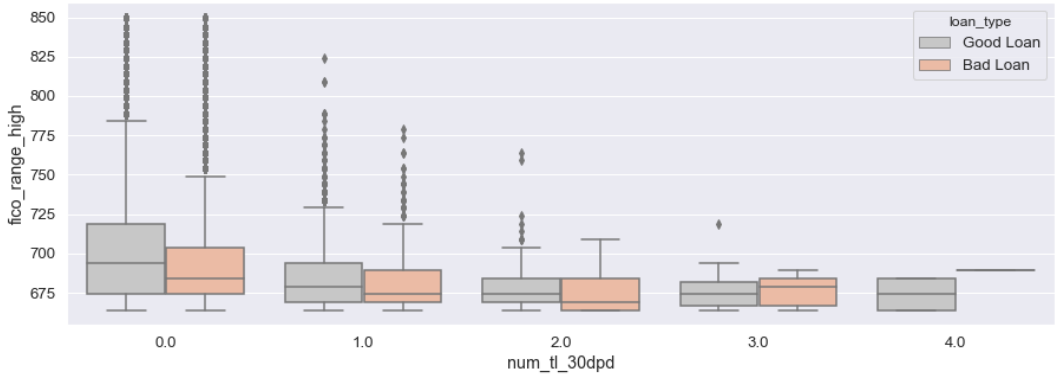
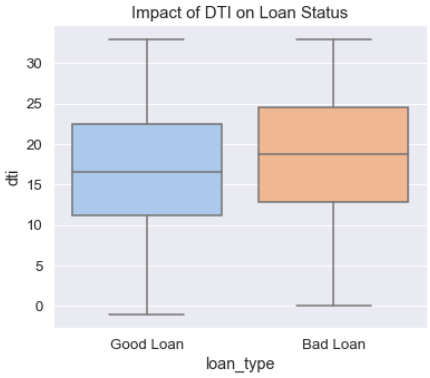


Fig 26

- **dti** (ratio using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income)



One of the important variables in the dataset that might impact the borrower's credit history is DTI, which gives the ratio between the borrower's total current monthly debt payment to his total reported monthly income. This can obviously impact his capability to repay the loan for which he is applying for.

Figures 27 and 28 show the impact of DTI on the loans as well as LC's grading system. They clearly show that the DTI is more for bad loans, and as we move towards riskier loans, the mean DTI increases.

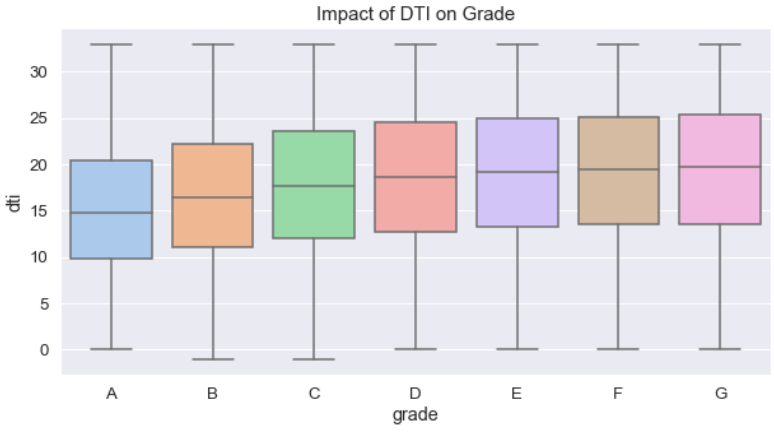


Fig 28

- **pub_rec_bankruptcies** (Number of public record bankruptcies)

The dataset gives us information on the number of public bankruptcies that each borrower has. From the graph in figure 29, we find that there are 3 unique values for the number of bankruptcies.

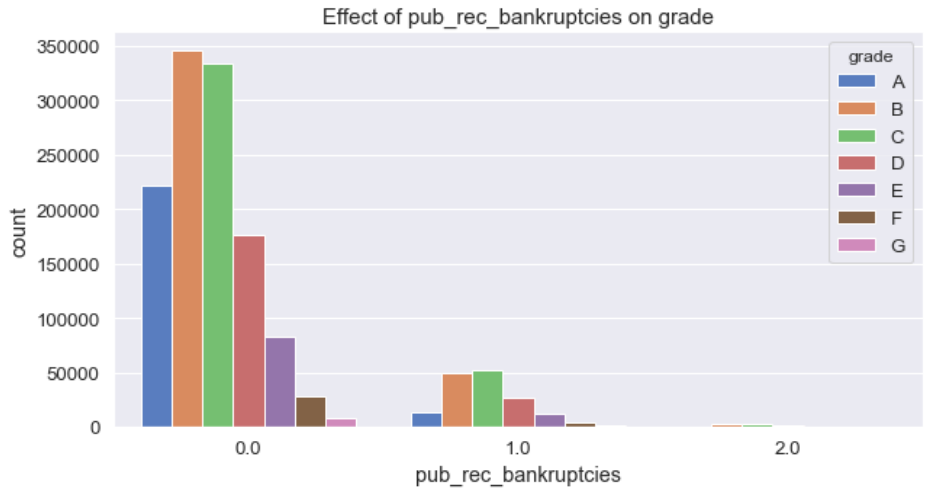


Fig 29

- Data Pre-processing

As part of data pre-processing and clean up steps, we have dropped redundant variables like member_id (unique LC assigned id for the borrower member), url (url for the LC page with listing data), zip_code (first 3 numbers of the zip code provided by the borrower in the loan application), etc and also variables having more than 50% missing values. The remaining attributes with missing values, have been treated with different imputation techniques.
Apart from that, other data cleaning techniques like outlier handling and variable transformation methods have also been applied. Finally, in order to be able to use various distance-based machine learning algorithms, we have scaled the variables using z-score technique.
As explained in the previous sections, the project has a 2-fold solution.

- Default Prediction Model

Our first data model is a default prediction problem, where we made an attempt to analyze the various attributes of the dataset and predict whether an ongoing loan will default or not. Thus, it forms a binary classification problem, where our final machine learning model needs to predict a loan as good or bad. For this, we have the target variable “loan_status” in the dataset.
We first select a sample from the dataset having only fully repaid loans and the ones that ended as default. On this sample, we perform various data cleaning and pre-processing steps to prepare it for the machine learning model. We observed that there is huge imbalance in the loans where we have a greater number of loans that ended successfully. Hence, we performed an oversampling technique to overcome this imbalance.
Next, we used various classification algorithms to assess their performance on the dataset.

- Loss Given Default Model

Our second model predicts the impact of a bad loan (Loss Given Default) on the lender. From the problem definition, we understand that this is a regression problem. For this, we used the attribute "recoveries" from the dataset. As per LC website, this variable defines all the recoveries that have been done post charge off. So, for model development and training purpose, we considered only those samples which had status as "Charged Off". In production, all the loans that are predicted as bad loans will be passed through the second model to see how much the lender can expect to lose if the loan actually defaults.

For the regression problem, we have used different regression algorithms.

Detailed Walkthrough of the Solution

- **Data Pre-processing**

Once the data was received, we found that it had lot of records which were summarizing the dataset and was not suitable for our modelling purpose. So, we dropped all such records.
There were few variables which did not make much sense and hence were dropped. Below are some such instances.

Feature Name	Feature Description	Reason for Dropping
acc_now_delinq	The number of accounts on which the borrower is now delinquent	Contains only a single value
collection_recovery_fee	post charge off collection fee	Considered a future variable as it will come into effect only the loan has charged off and recovery has been made
pymnt_plan	Indicates if a payment plan has been put in place for the loan	Contains only a single value
hardship_type	Describes the hardship plan offering	Only 0.01% of the samples were of type "hardship loans". Hence, we didn't consider these variables.
url	URL for the LC page with listing data	Unique id field
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application	Didn't provide any information

Table 2

Next, we looked at the missing content of each variable. Any variable having more than 50% values missing were dropped. For the rest, depending on the description of the variable, we have tried different imputation techniques like Zero Imputation, Median and Mode imputation as well as "Missing" indicator imputation techniques.
Imputation was performed keeping in mind that the process should not heavily distort the original distortion. Few examples are given below.

- Before Imputation:

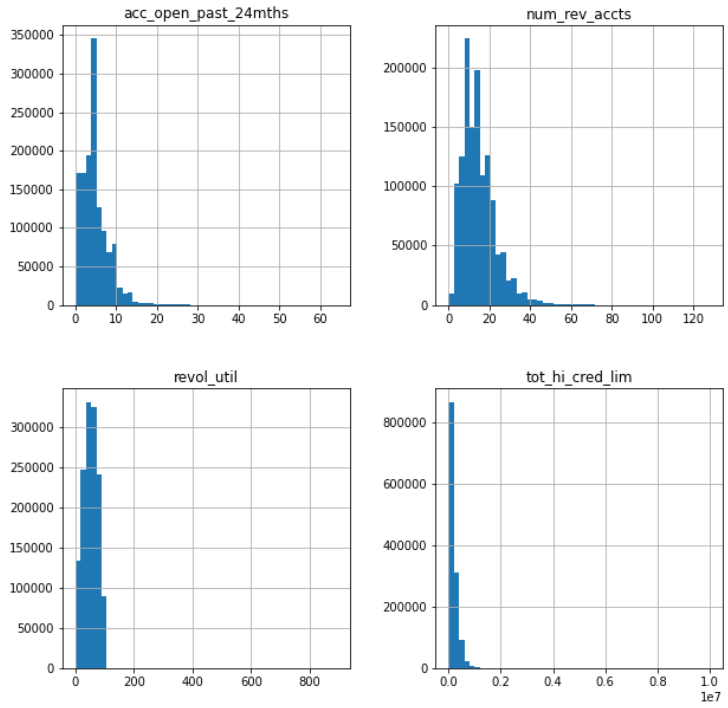


Fig 30

- After Imputation:

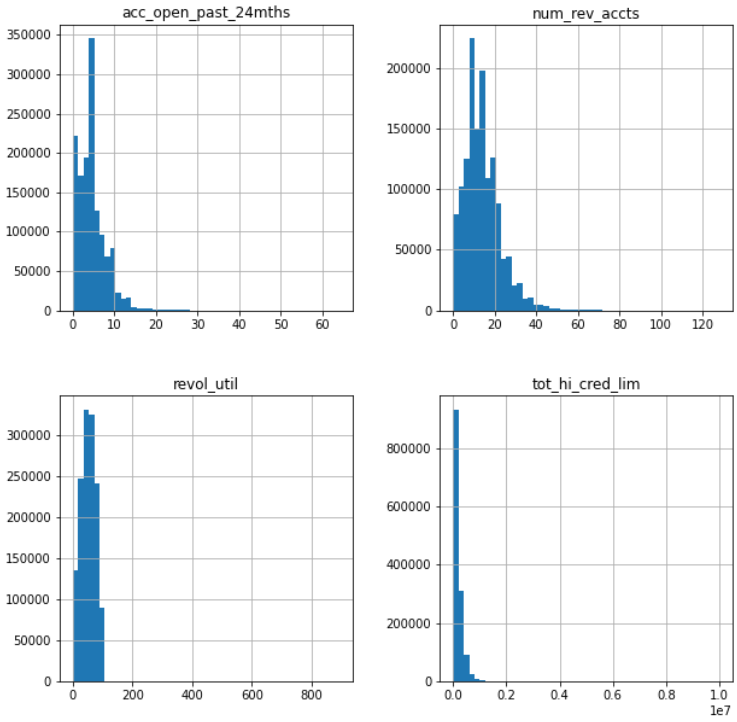


Fig 31

In our regression problem, after missing value imputation, we looked at normalizing the distributions of the different features. For this, we used log, power, yeo-johnson and other transformation techniques. Some instances of the transformation process are given below in figures 32 and 33.

- annual_inc transformed using Power Transformation

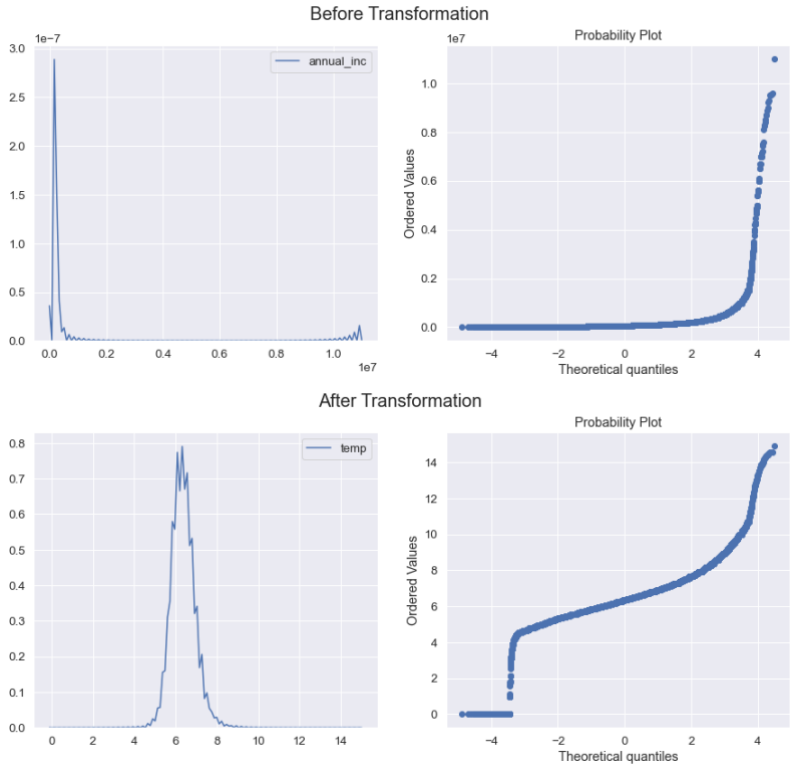


Fig 32

- int_rate transformed using log Transformation

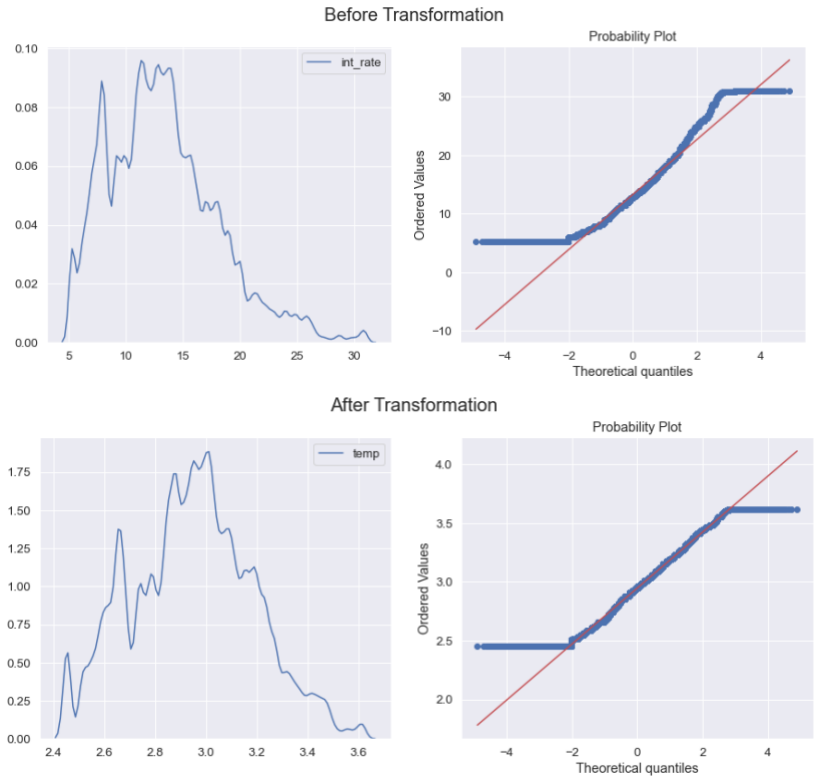


Fig 33

We find that after the transformation is done the skewed distributions tend to become a normal one. Also, the q-q plot shows that the points somewhat follow a straight line after the transformation is done.
After the imputation process (in case of classification only) and distribution transformation, we looked at the categorical variables like “term”, “sub_grade”, “home_ownership”, etc which were converted into numerical variables.
In order to be able to use distance-based machine learning algorithms, we have standardized the features of the dataset, i.e, brought the ranges of the features to a similar scale. This is done by putting the mean of the distribution at 0 and having a unit standard deviation. For this, we use the formula:

$$X' = \frac{X - \mu}{\sigma}$$

Where,
X’ = standardized value
X = original value
μ = mean of the feature values
σ = SD of the feature values

Below (figures 34 and 35) are examples of few features before and after the standardization process. If we compare the ranges of the features, we find that after standardization, the ranges of the distributions are somewhat comparable with their means around 0.

Before Standardization:

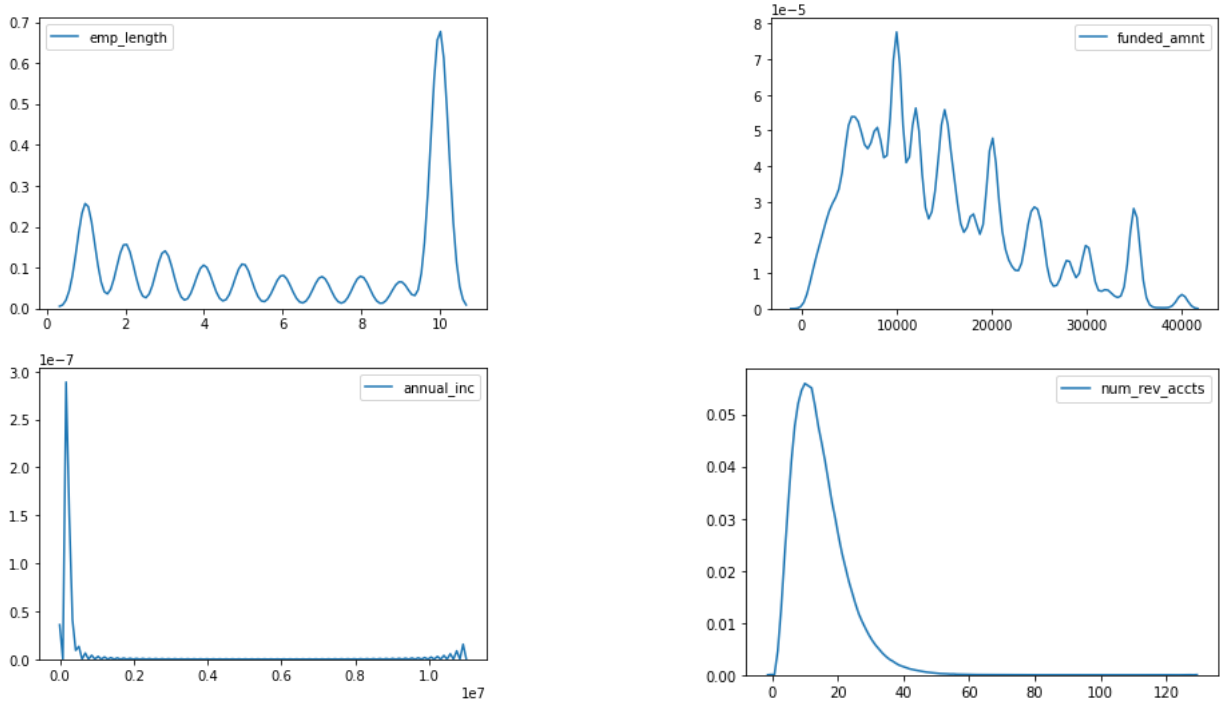


Fig 34

After Standardization:

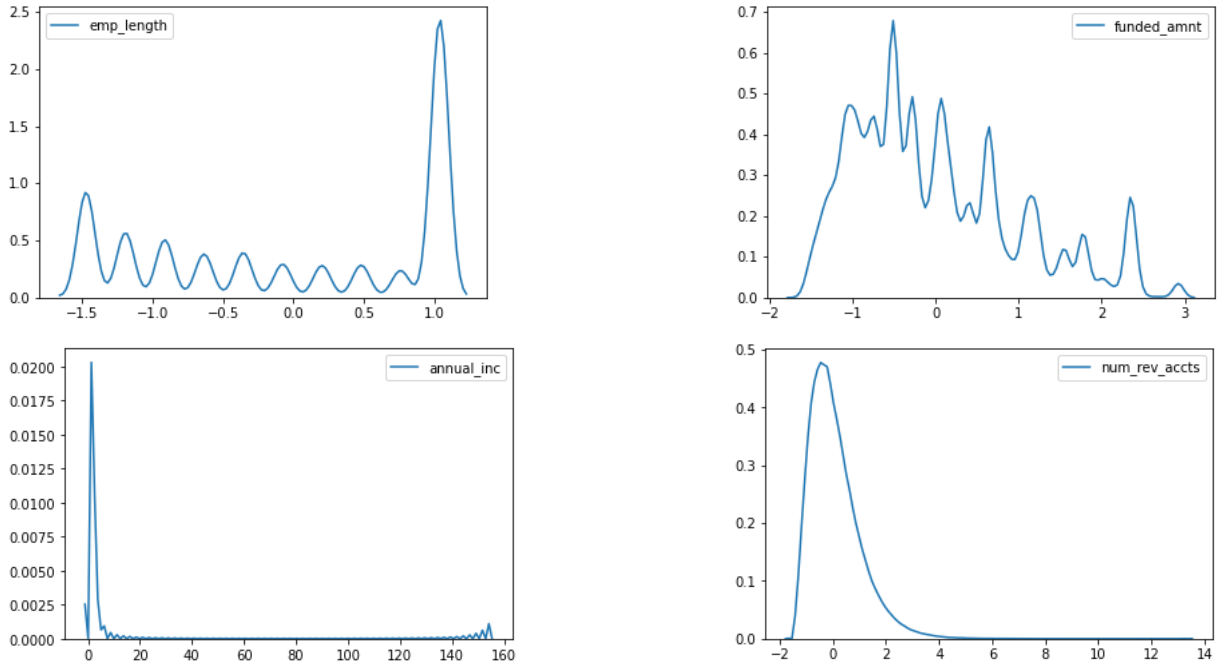


Fig 35

Although, as shown earlier, we observed high linear relationship between few of the variables and decided to drop them. Here, we try to explore more into the multi-collinearity between the variables through the heat map in Figure 1 at the end of the report.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

We have plotted the heatmap by calculating the Pearson's correlation between the different variables.

- **Classification Problem**

For our classification problem, the target variable was "loan_status". As mentioned earlier, the feature had values as:

1. Current: Active Loan.
2. Fully Paid: The full principal with interest rates is paid back.
3. Charged Off: The borrower defaulted on the loan and the loan will never be paid back in full amount.
4. In Grace Period: Payment of installment is delayed by 1 to 15 days.
5. Late (16–30 days): Payment of installment is delayed by 16 to 30 days.
6. Late (31–120 days): Payment of installment is delayed by 31 to 120 days.
7. Default: Payment of installment is delayed by more than 120 days.

We considered a loan with status **Fully Paid** as "**Good Loan"** and the one having status **Default** or **Charged Off** as **"Bad Loan"**. Lending Club has observed that borrowers who do not pay their

due installment for more than 30 days, have ultimately defaulted on their loans. Hence, we considered loans with status **Late (31–120 days)** as "Bad Loan". This left us with loans **Current**, **In Grace Period** and **Late (16–30 days)**, which we considered as **"Current Loan"** since we cannot exactly say how these loans might end. Thus, we map the different statuses in the dataset as,

- Good loan: Fully Paid
- Bad Loan: Charged Off, Late (31–120 days), Default
- Current Loan: Current, In Grace Period, Late (16–30 days)

For our problem at hand, we considered only the loans with status as **Good** or **Bad** to build our model and then used the Current loans to predict which of these active loans are going to default. *Hence, we kept our Current loans aside for the time being and focused only on the Good and Bad loans.*

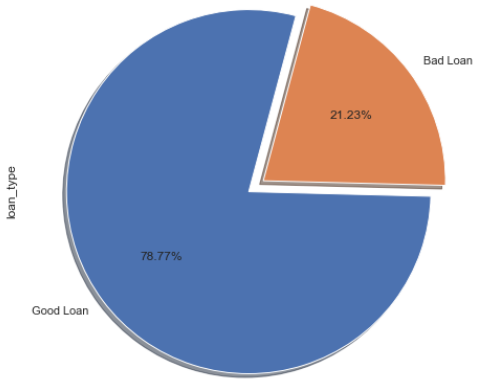


Fig 36

After the above-mentioned operation, we found that there was a total of 1369566 records for Good and Bad loans of which approx 79% are marked as Good Loan and 21.33% are marked as Bad Loan. (figure 36)
It was quite evident that there was a huge difference in the total number of good and bad loans. So, with this data in hand, whatever model we built would always have been bias towards the Good Loans. Hence, we needed some oversampling technique to remove this imbalance from our dataset.
Two possible over sampling techniques are Random Oversampling and Synthetic Minority Oversampling Technique (SMOTE). SMOTE creates a new sample by observing the closest samples in the feature space. As a result, the performance of SMOTE is too low. Hence, we decided to go with Random Oversampling technique which although just duplicates the minority class, gave us quick results. Also, we compared took 10% of the dataset and compared the performance of both SMOTE and Random OverSampler and found the performance to be more or less the same.
After oversampling, we found our dataset comprising of 2157478 samples.
Once, the dataset was ready, we divided the data keeping 15% for testing and remaining 85% for model training purpose and tried the different classification problems comparing their ROC curves and AUC scores.

- **Regression Problem**

For building a machine learning model to determine the loss that the lender has to bear in the event of a default, we picked only those samples from the dataset with final status as **"Charged Off"**. We found that to be a total of 2,68,559 records. As discussed in the previous sections, the target variable in this problem was "recoveries".

[13] defines LGD as,

$$LGD = 1 - \frac{total\_payment}{total\_amount}$$

where, total_payment = total loan amount that has been recovered so far, and

total_amount = total loan amount that should have been recovered.

Ideally, we should factor in the fees for recovery as well. Hence, the equation should have been,

$$LGD = 1 - \frac{total\_payment - \ collection\_fees}{total\_amount}$$

However, it is difficult to predict the collection fees until the recovery operation has happened. Hence, we considered only the former equation for calculating LGD.

We also tried to determine the optimal number of clusters that can be formed from this new dataset. For this, we used K-Means Clustering technique and both the Elbow as well as the Silhouette Analysis techniques to analyse the results.

- Elbow Method:

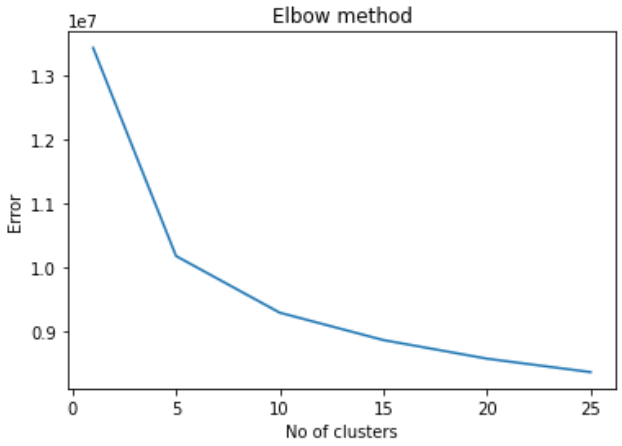


Fig 37

From the figure (), we find that there is a sharp elbow forming with 5 clusters and faint cluster forming with 10 clusters. However, even after that, we find the squared error reducing significantly. Hence, we tried Silhouette Analysis to determine the most optimal number of clusters.

- Silhouette Analysis

Continuing with our findings from the Elbow Method, we tried with 5, 10, 25 and 30 clusters and below are the results (figure 38).

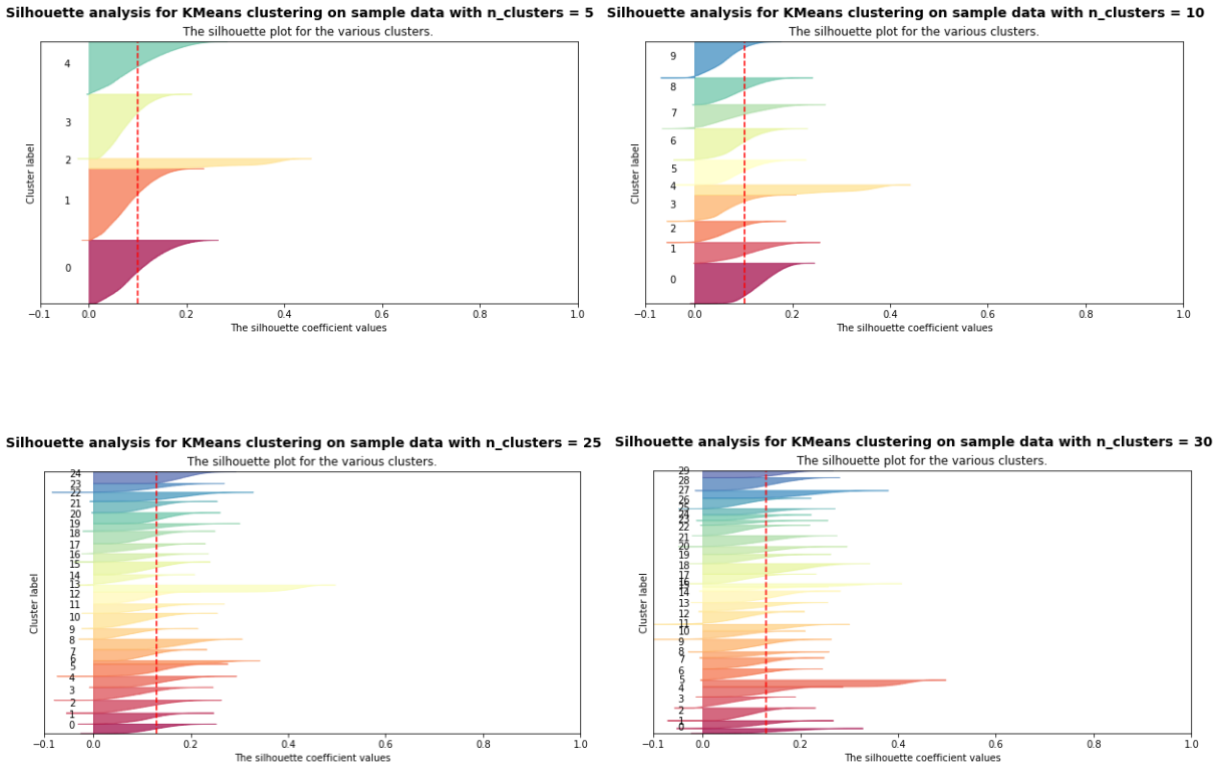


Fig 38

Number of Clusters	Silhouette Score
5	0.1006
10	0.1022
25	0.1306
30	0.1298

From the above results, we find that 5 clusters seem to be an optimal value. However, from the silhouette diagrams, we see that there is a little overlap between the clusters.

Once, the dataset was ready, we divided the data keeping 15% for testing and remaining 85% for model training purpose and tried the different regression problems comparing their Adjusted R-Squared values.

Model Evaluation

- Default Prediction – Model Selection

As mentioned earlier, we used several classification models for default prediction and compared their performances, which is summarized in the below table.

Model Name	Hyperparameter Values
Logistic Regression	solver='sag' C=1 penalty='l2' multi_class='auto' max_iter=10000
Decision Tree	criterion='entropy' min_samples_leaf=1 min_samples_split=2 max_depth=9 max_features='auto' max_leaf_nodes=20
Random Forest	criterion='entropy' min_samples_leaf=1 min_samples_split=2 max_depth=9 max_features='auto' max_leaf_nodes=21 n_estimators=400
K-Nearest Neighbors	n_neighbors=22 algorithm='auto' weights='distance' p=1
Support Vector Classifier	kernel='rbf' gamma='auto' class_weight='balanced' C=1000 max_iter=10000

Table 3

Also, below is a table of the performance metrics of the different models. We compared the AUC scores of the models to select the best one.

Model Name	Accuracy	Precision	Recall	F1	AUC Score
Logistic Regression	0.9939	0.9925	0.9998	0.9961	0.999
Support Vector Classifier	0.9922	0.9918	0.9983	0.995	0.998
Random Forest	0.9446	0.9532	0.9356	0.9443	0.986
K Nearest Neighbors	0.9032	0.8943	0.9942	0.9416	0.964
Decision Tree	0.8895	0.923	0.8512	0.8856	0.915

Table 4

From the AUC scores in table 4, we find Logistic Regression to give the best performance and hence, we selected Logistic Regression as our final algorithm for classifying good and bad loans.

- Default Prediction – Model Description

Logistic Regression (LR) is defined as a statistical model that uses a logistic function to model a binary classification problem. Here, we are using it to classify between Good Loans which is numerically labelled as 1 and Bad Loans numerically labelled as 0.

Our LR model used an L2 approach for regularization in order to avoid overfitting. So, since we used L2 and also since we had a somewhat large dataset in our hand, we used "sag" as our solver. We evaluated our model's performance on different metrics like accuracy, precision, recall, and f1.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The accuracy of our model is 0.9939. This means that during testing about 99.39% of the predictions made by our classification model is correct.

In order to understand how the model performed for each class, we have created the confusion matrix for the model (as given in figure 39). From this we can see that about approx. 98% of the bad loans were predicted correctly during testing of the model. Similarly, about 99.9% of the good loans are predicted correctly. Thus, we can say that we have built a pretty good classifier, which can predict both the good and bad loans equally well.

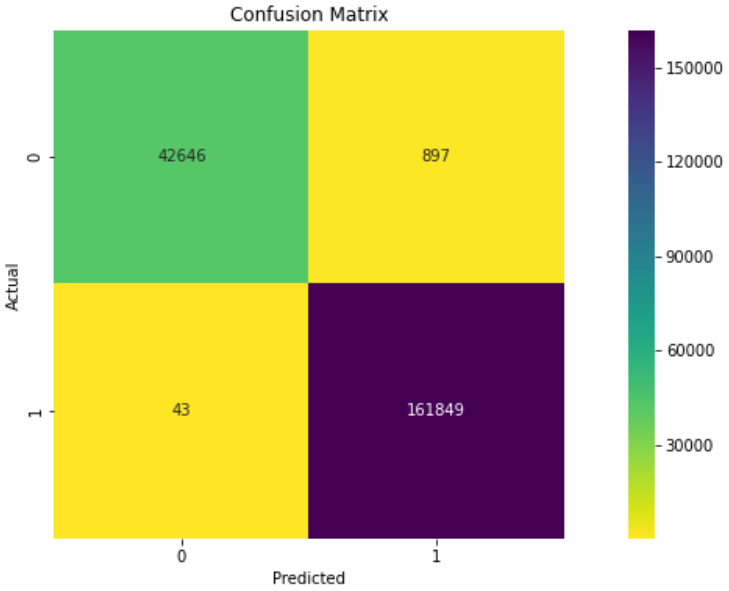


Fig 39

Further, in order to confirm the performance of our classification model, we look at its ROC score which comes at 0.999 (AUC curve in figure 40). A value so close to 1 means that we have an excellent classifier in place.

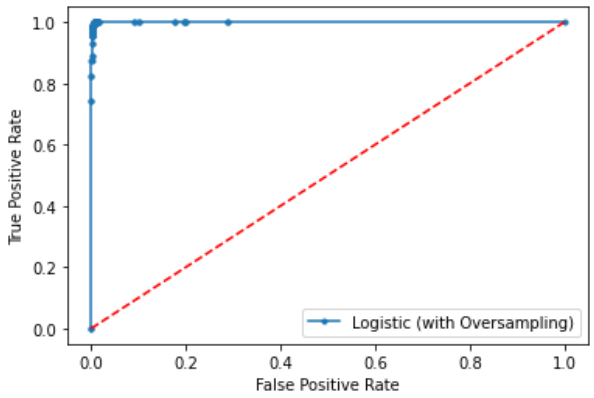


Fig 40

We also looked at the important features in the model. Below is a table of the top 5 features deciding whether a loan will end up as a Good Loan or a Bad Loan.

Feature Name	Feature Importance (for Good Loans)	Feature Name	Feature Importance (for Bad Loans)
total_rec_prncp	22.765821	funded_amnt	-20.220704
last_fico_range_high	1.526253	total_rec_int	-0.991154
term	0.501807	hardship_flag	-0.217384
last_pymnt_d	0.378949	total_rec_late_fee	-0.177522
issue_d	0.220179	total_acc	-0.128451

Table 5

A detailed list of feature importance is given in Table 1 at the end of this report.
We find that the above listed features play the most important role in deciding the final status of a loan. In that, total_rec_prncp and funded_amnt is heavily dominating the other features. This created some suspicion that the features might not be valid for our scenario.
As per definition given by lending Club,

- total_rec_prncp - Principal received to date
- funded_amnt - The total amount committed to that loan at that point in time

Since, we are trying to predict how the currently active loans of lending Club will end up, we need to know the loan amount as well as the principal received for the loan till date. Hence, these two features are very important for default prediction and we decided to keep them in our model.

- Loss Given Default – Model Selection

As mentioned earlier, we used several classification models for default prediction and compared their performances, which is summarized in the below table.

Model Name	RMSE	R^2	Adjusted R^2
Linear Regression	1094.03	0.2801	0.2789
Polynomial Regression	1037.18	0.35	0.3117
Ridge Regression	1051.72	0.2801	0.2789
Lasso Regression	1044.01	0.2799	0.2787
Random Forest Regressor	977.44	0.4254	0.4244

Table 6

We also tried to implement Artificial Neural Network, but found very low scores compared to the Normal Machine Learning algorithms.
From the above table, we find that Random Forest Regressor produces the best results. Hence, we select it as our final model for LGD prediction.

- Loss Given Default – Model Description

Random Forest Regressor is an ensemble technique that performs regression using multiple decision trees. The basic idea behind this is to combine multiple decision trees in determining the

final output rather than relying on individual decision trees. In this algorithm, we perform random row and column sampling from the dataset forming sample datasets for every model.
For our problem, we found the below hyper parameters to be the most optimal one.

Parameter Name	Parameter Value
n_estimators	200
min_samples_leaf	1
min_samples_split	2
criterion	Mean Squared Error
max_depth	9
max_leaf_nodes	21

Table 7

We know that R-Squared determines the percentage of variance in the dependent variable, that the independent variables can explain. Here, our R-Squared value is 0.42 or 42%. That means that only 42% of the variation can be explained by our model.
Since adjusted R-Squared considers only those variables which contribute significantly to our model, we considered the adjusted R-Squared value over the former one. Here, our adjust R-Squared value is almost the same as the R-Squared value.
RMSE is a measure of how spread out these residuals are. In other words, it tells us how concentrated the data is around the line of best fit. Our RMSE score is 977.44. Although compared to other algorithms it is less, still the value is quite high, stating that our model has not done a very good job in predicting the loss that a lender would incur in the event of a default.
One of the primary reasons why our regression model is not able to produce a good result is the skewness that is there in the dependent variable. Because of this skewness (shown in figure 41), our model is highly lenient towards values closer to 0.

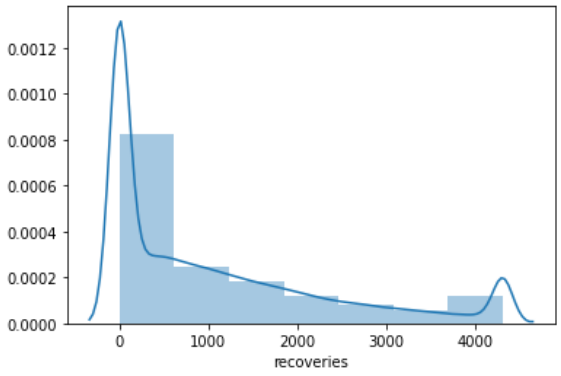


Fig 41

Benchmark Comparison

At the onset of our project, we used Logistic Regression as our base model with an AUC score of 0.72 (figure 42).

We need to note that, while building our base model, the EDA process was going in parallel and hence not all dataset features were included in the model. Also, the features did not undergo any normalization and standardization techniques.

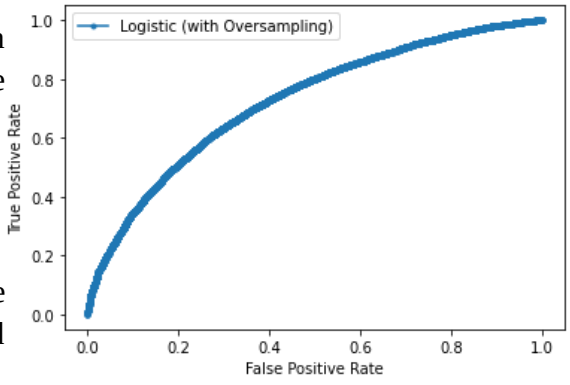


Fig 42

Now, with all feature engineering techniques complete and all the required features included, when we re-trained our model, we found the performance has considerably improved to an AUC score of 0.99.

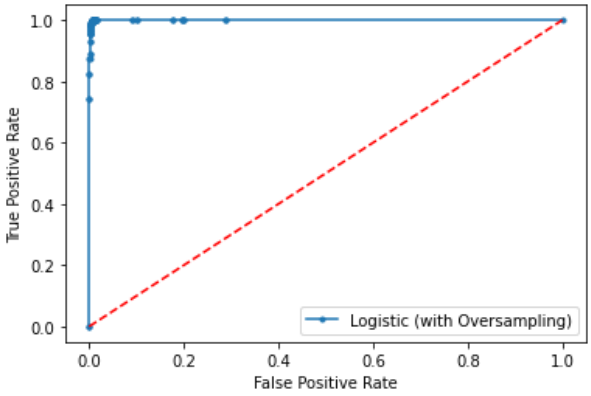


Fig 43

In previous research works done like in [21], we find that the best performance was an AUC score of 0.983 (using Random Forest Classifier). Here, we have exceeded that value and can confidently say that we have built an excellent classifier for default prediction.

For our regression problem, we didn't find much research work done on recoveries prediction and hence this stays as somewhat of a future research work for us.

Implications

Default Prediction is one of the most important fields of research for any lending institution in today's market. Our solution in this project will help lenders of Lending Club understand how the current loans that they have funded will turn out. If they find that the loans that they have funded has a probability to default, then our solution will also give them a rough estimate of how much money is at risk.
Also, our findings regarding the determinants of a loan default event will help the lenders select prospective loan applications in the future.

Limitations

The main limitation in our model is the low accuracy with which it is predicting the loss that a lender will incur in the event of loan default. In real world, sometimes lenders are okay even after knowing the loans that they are funding might default since they expect higher returns. Such low levels of accuracy might be risky for these lenders as in some of these predictions we have seen that the amount at risk was quite high, but the model predicted low amount at risk.
We already analysed that the main reason for such bad performance by our regression model was due to the skewness that existed in the dependent variable. In order to enhance this, we need to look at ways to remove this skewness.
In [22] we have seen an approach to perform regression for imbalanced dependent variable through an algorithm called SMOGN and this is something we need to look into in order to improve our regression performance.

Also, we need to look into ways to introduce distributed execution through cloud resources while training our model so that we can implement algorithms that were not feasible with our current infrastructure.

Closing Reflections

Through this project, we have found out that attributes like the funded loan amount, total principal and interest received so far loan tenure, lendee's FICO score play vital roles in determining whether the loan will eventually default or not.

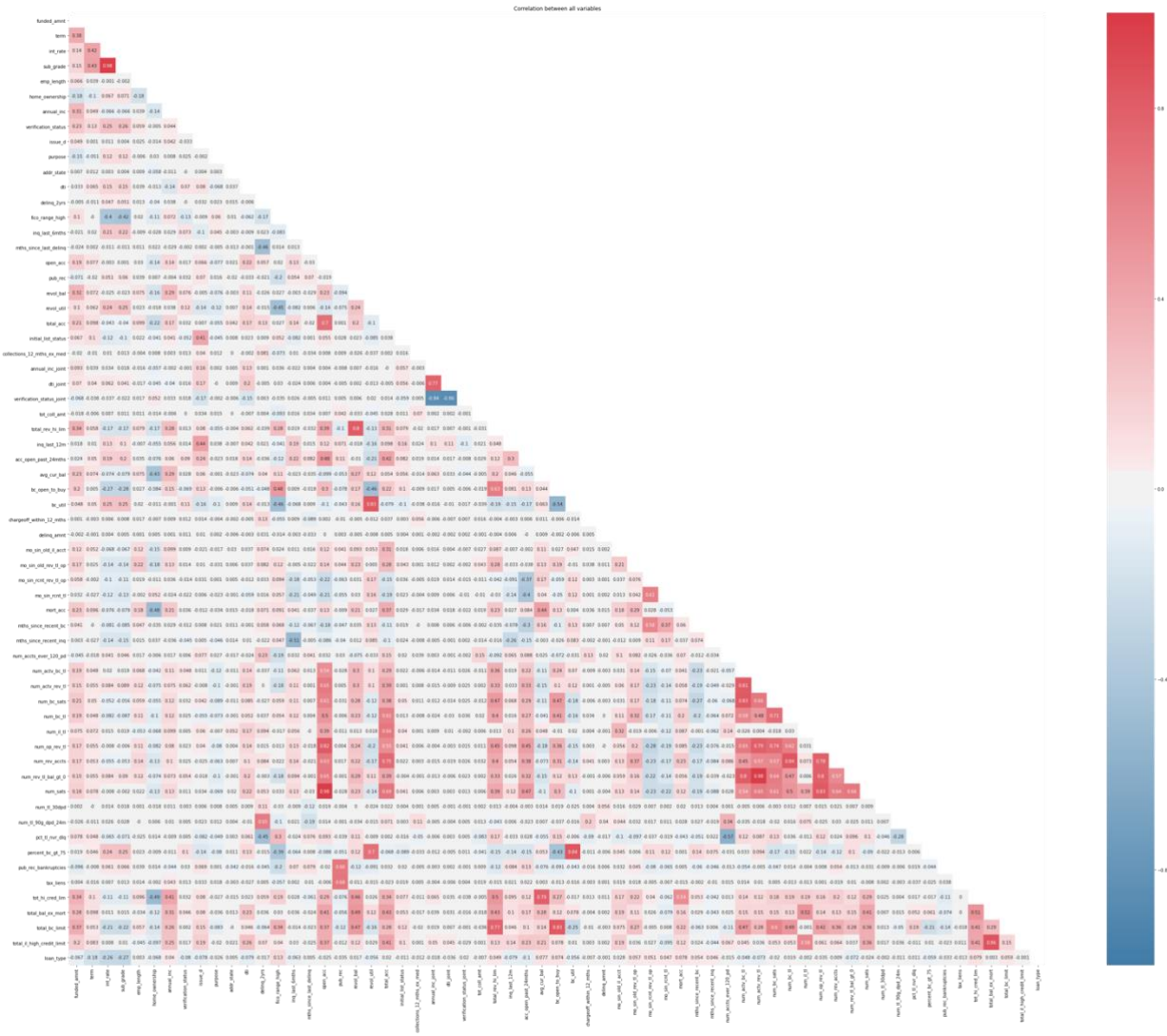
Also, with the data that we have Logistic Regression is the best classification model for Default Prediction. Also, with the same data, Random Forest Regression can be used to determine Loss Given Default (how much the lender is expected to lose in the event of a default).

In order to improve our model further, we would like to look at algorithms that take longer execution time by employing distributed execution.

Table 1

Features deciding Bad Loans	Feature Importance	Feature deciding Good Loans	Feature Importance
funded_amnt	-20.220704	initial_list_status	0.001824
total_rec_int	-0.991154	mths_since_recent_bc	0.003452
hardship_flag	-0.217384	pub_rec_bankruptcies	0.007054
total_rec_late_fee	-0.177522	mths_since_last_delinq	0.008429
total_acc	-0.128451	tax_liens	0.008641
bc_open_to_buy	-0.108019	mo_sin_old_il_acct	0.011208
mo_sin_old_rev_tl_op	-0.104005	num_sats	0.017962
num_bc_tl	-0.073505	mort_acc	0.020196
bc_util	-0.071255	annual_inc	0.021126
last_credit_pull_d	-0.064792	avg_cur_bal	0.029781
fico_range_high	-0.064197	dti_joint	0.031385
open_acc	-0.061272	num_accts_ever_120_pd	0.036386
dti	-0.051507	num_il_tl	0.039755
emp_length	-0.046502	verification_status_joint	0.040833
num_op_rev_tl	-0.042417	acc_open_past_24mths	0.046516
num_tl_90g_dpd_24m	-0.034761	percent_bc_gt_75	0.050051
verification_status	-0.033956	revol_util	0.058122
total_bal_ex_mort	-0.031669	delinq_2yrs	0.060393
num_actv_bc_tl	-0.027890	pct_tl_nvr_dlq	0.063277
home_ownership	-0.027810	total_il_high_credit_limit	0.087392
inq_last_12m	-0.026653	num_rev_accts	0.100289
pub_rec	-0.023027	num_bc_sats	0.123038
annual_inc_joint	-0.021780	sub_grade	0.169488
revol_bal	-0.020074	issue_d	0.220179
purpose	-0.019149	last_pymnt_d	0.378949
mths_since_recent_inq	-0.017456	term	0.501807
total_bc_limit	-0.014192	last_fico_range_high	1.526253
addr_state	-0.013939	total_rec_prncp	22.765821
tot_hi_cred_lim	-0.011188		
collections_12_mths_ex_med	-0.010428		
num_actv_rev_tl	-0.008517		
inq_last_6mths	-0.007061		
total_rev_hi_lim	-0.005645		
chargeoff_within_12_mths	-0.003018		
mo_sin_rcnt_tl	-0.002317		
mo_sin_rcnt_rev_tl_op	-0.001823		
num_tl_30dpd	-0.001146		
delinq_amnt	-0.001125		
tot_coll_amt	-0.000675		

Figure 1



References

[1] B. Senthil Arasu, P. Sridevi, P. Nageswari, R. Ramya, "A Study on Analysis of Non-Performing Assets and its Impact on Profitability", International Journal of Scientific Research in Multidisciplinary Studies, Volume-5, Issue-6, pp.01-10, June (2019) (d)
[2] Yang Yang, "An empirical study on P2P loan default prediction model", Financial Engineering and Risk Management (2020) 3: 14-22, Clausius Scientific Press, Canada
[3] Michal Polena, Tobias Regner, "Determinants of Borrowers' Default in P2P Lending under Consideration of the Loan Risk Class", Games 2018, 9, 82 (d)
[4] Rajkamal Iyer, Asim Ijaz Khwaja Erzo, F. P. Luttmer, Kelly Shue, "Screening in New Credit Markets: Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending?", SSRN-id1570115 (d)
[5] Don Carmichael, "Modeling Default for Peer-to-Peer Loans", SSRN-id2529240
[6] Alan Zhang, "DEVELOPMENT OF LOGISTIC REGRESSION MODEL TO PREDICT DEFAULT PROBABILITIES OF LOAN ACCOUNTS", International Journal of Information, Business and Management, Vol. 12, No.2, 2020, pg 95-115 (d)
[7] Anand Motwani, Goldi Bajaj, Sushila Mohane, "Predictive Modelling for Credit Risk Detection using Ensemble Method", International Journal of Computer Sciences and Engineering, June 2018, Vol-6, Issue-6 (d)
[8] Zakaria Alomari, Dmitriy Fingerman, "Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications", New Zealand Journal of Computer-Human Interaction ZJCHI 2,2 (2017)
[9] Pedro G. Fonseca and Hugo D. Lopes, "Calibration of Machine Learning Classifiers for Probability of Default Modelling", James Finance (CrowdProcess Inc.), October 24th, 2017 (d)
[10] Rising Odegua, "Predicting Bank Loan Default with Extreme Gradient Boosting" (d)
[11] Jose A. Lopez, Marc R. Saidenberg, "Evaluating Credit Risk Models", Journal of Banking & Finance, Volume 24, Issues 1-2, January 2000, pgs:151-165
[12] Jeremy Turiel, Tomaso Aste, "P2P Loan Acceptance and Default Prediction with Artificial Intelligence", https://www.researchgate.net/publication/334223307
[13] Guangyou Zhou, Yijia Zhang, Sumei Luo, "P2P Network Lending, Loss Given Default and Credit Risks" (d)
[14] Felix Martinson, "Exotic Approaches for Modelling Loss Given Default" (d)
[15] Shunpo Chang, Simon Dae-Oong Kim, Genki Kondo, "Predicting Default Risk of Lending Club Loans"
[16] Carlos Eduardo Canfield Rivera, "Determinants of Default in P2P Lending: The Mexican Case", Independent Journal of Management & Production (IJM&P), v.9, n.1, January - March 2018
[17] Christophe Hurliny, Jérémy Leymariez, Antoine Patinx, "Loss functions for Loss Given Default Model Comparison"
[18] Peter Martey Addo, Dominique Guegan, Bertrand Hassani, "Credit Risk Analysis Using Machine and Deep Learning Models", Risks 2018, 6, 38
[19] Til Schuermann, "What Do We Know About Loss Given Default?", Forthcoming in D. Shimko (ed.), Credit Risk Models and Management 2nd Edition, London, UK: Risk Books, February 2004
[20] Han Sheng Sun and Zi Jin, "Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default", Journal of Credit Risk 12(3), pgs 43–69
[21] Lin Zhu, Dafeng Qiu, Daji Ergua, Cai Ying, Kuiyi Liu, "A study on predicting loan default based on the random forest algorithm"
[22] Paula Branco, Lu´ıs Torgo, Rita P. Ribeiro, "SMOGN: A Pre-processing Approach for Imbalanced Regression"